Quality Evaluation of Audio and Video Signals in Videoconferences

Jana Filanová, Iveta Ondrášová, Anikó Töröková

University of Economics in Bratislava
Dolnozemská cesta 1, 852 35 Bratislava, Slovak Republic
jana.filanova@euba.sk, iveta.ondrasova@euba.sk, aniko.torokova@euba.sk

Abstract: Videoconferencing represents a technology of the future, in modern education. A combination of audio and video information serves in understanding the content of lectures or presentations, in the form of videoconferencing. The evaluation of the quality of videoconferencing is difficult, as the image and sound affects the final quality. In general, occasional image disturbance has less impact on the perception of quality in comparison to the disturbances in an audio track. In this research, we simulated a real packet network environment and tested video sequences that present different teaching content. We artificially degraded the quality of video sequences by packet loss and jitter. Our test aimed to compare subjective methods of video quality evaluation with objective methods and to evaluate the impact of audio quality on the overall video sequence quality. This paper describes a novel process of evaluating the quality of audio and video signals. Time-consuming subjective measurements were supported by models and programs that simplified the preparation, testing, and processing of results. The contribution of this article is to present and evaluate the results of video sequence quality testing with an emphasis on semantics, which has a significant impact on viewers' sensitivity to video sequence quality.

Keywords: videoconferencing; virtual reality; quality evaluation of video and audio; packet loss; latency; objective assessment; subjective assessment; MOS scores; semantics

1 Introduction

Videoconferences represent a form of synchronous communication based on audio and video transmission with the possibility to integrate text and other forms of presentation of information at a distance. The quality of this communication is influenced by the used communication technologies and transmission characteristics of communication networks [1]. Videoconferencing is one of the most appropriate ways of online transmission information to participants. The videoconferences could be recorded and it is possible to view the records even in the off-line mode [2].

At the primary and secondary education levels, videoconferencing can be used for teaching pupils without access to regular education. This might be due to the students' physical isolation (e.g., students living in remote areas, disabled students or students quarantining at home) or for various economic and social reasons. In addition, videoconferencing can be applied to the teaching of gifted students who can benefit from more intense learning or choice of subjects not available at their school. Teaching and learning are complex activities realized by many methods [3].

Universities, High Schools, and various Higher Education Institutions are trying to meet the needs of growing numbers of external students, whose, other commitments, do not allow them to attend regular lectures and exercises. The modern trend in education is virtual reality. It represents a modern form of education, which brings education content from the classical education room to an online environment. Students and teachers have then access to education and information from anywhere [4].

The visual perception of people is a highly complex matter that involves several mechanisms. It is influenced by their expectations and their previous experience. The view of the quality is linked to their mechanisms of imagination. The quality of the presentation through videoconferencing will depend not only on the technical quality of the videoconferencing but also on other factors such as lecture content [5]. In [6] the authors show that semantics has a significant impact on viewers' sensitivity to the quality of a video sequence for spatially separated parts of the sequence and, more importantly, that this difference in sensitivity can be changed by the presence of an audio signal. This result is important for any testing of subjects' responses to visual material. One example is the subjective assessment of the quality of video in an audio-visual communications system (such as television or videoconferencing) [6].

Videoconferencing quality testing is very specific. In the real world, we usually perceive information simultaneously from two or more sources and then process them into the resulting form. A good example is the reading from lips where, besides the speaker's voice, we also observe the movement of his/her lips. From the perspective of subjective evaluation of videoconferencing quality, it is true that some parts captured by the camera are more important than others. Such areas are known as "Foregrounds". For example, during a videoconferencing, the most important areas are the head and shoulders of the person being captured, while the rest in the background is not important [7].

The human eye is the most important organ in sensory perception. Human beings acquires about 80% of the world's information using their eyes. But one must realize that the eye does not give the brain a definite picture of the outside world. The image of the outside world consists of a combination of information from the eye and the observer's experience [8]. The transition from the stimulus in the eye to the central nervous system analysis is not immediate but has a delay of

approximately 20 ms (on average and differs from person to person). This means that patterns changing at a rate greater than 50 Hz are perceived as continuous movements [9]. For instance, the television works on the same principle.

The sound is defined as every longitudinal mechanical oscillation in a medium that is capable of creating a hearing perception in the human ear. The sensitivity threshold of the auditory organ in a healthy human is about I0 = 10-12 Wm-2 at a frequency of 1000 Hz. This amount is referred to as the zero volume level or the conventional listening threshold (0 dB) at a frequency of 1000 Hz [10]. At this threshold sound intensity, the amplitudes of the movement of the eardrum are of the order of the atom diameter. The basilar membrane oscillations show approximately the same amplitudes. According to current knowledge, it is difficult to explain the mechanism by which these slight deflections can cause irritation of the nerve endings [11].

The results of the research [12] have shown that the presence or absence of audio has a significant impact on the overall subjective perception of the videoconferencing quality. It has also been found that the viewer is more sensitive to the quality of the image in the foreground of the speaking person than to the quality of the image in the background. If there are multiple people in the scene, even not speaking right now, the viewer is likewise more sensitive to the quality of the image of the captured people than to the quality of the image in the background [12].

Digital image data stored in image databases and distributed over communication networks are subject to various types of distortions during data acquisition, compression, processing, transmission, and reproduction. e.g., lossy video compression methods that are almost always used to reduce the bandwidth needed to store or transmit video data may degrade video quality during the quantization process. In fact, digital video streams transmitted over error-prone channels (e.g., wireless channels) may be received as incomplete due to the deterioration encountered during the transmission. Packet communication channels (Internet) can cause loss or delay of received packets, depending on network status and QoS (Quality of Service) used [2]. The effects of time delay can be reduced with various control methods designed for latency-tolerance [13]. Transmission errors can result in a deterioration of the received image information. Therefore, it is desired that systems designed for video services are able to realize and quantify the degradation of video quality that occurs in the system. This is especially important in order to maintain, manage, and at best, improve the quality of image data. Effective metrics of quality of static image and video are essential for this purpose [14].

Image quality assessment is a challenging task that is traditionally approached by computational models. To maintain, control, and enhance the quality of images, it is important for image acquisition, management, communication, and processing systems to be able to identify and quantify image quality degradations. A great

deal of effort has been made in recent years to develop objective image quality metrics that correlate with perceived quality measurement [15, 16].

The aim of developing new methods for evaluating video quality objectively is to design metrics that can independently predict video quality [15]. Objective video metrics can be used to monitor image quality in quality management systems. When using objective video metrics, a network video server can monitor the quality of video transmitted by the network and manage video streaming. Objective video quality measures play important roles in various video processing applications, such as compression, communication, printing, analysis, registration, restoration, and enhancement. Experiments on the video quality experts group (VQEG) test dataset show that the new quality measure has a higher correlation with subjective quality measurement than the proposed methods in VQEG's Phase I tests for full-reference video quality assessment [17].

The most reliable way to measure video quality is subjective assessment because in most cases, a human being is the ultimate recipient of the video.

However, one of the major issues is that subjective methods are inconvenient, slow, and costly for practical use.

This article presents the process of quality evaluation of video sequences, which gives practical instructions to facilitate and accelerate subjective evaluation. Section 2 explains the methodology of our research. It describes subjective and objective methods for evaluating video and audio and finally a process model used in our research. Section 3 presents the results of the research including a comparison of the video sequences quality evaluation results. The contribution of the article is to present and evaluate the results of video sequence quality testing with an emphasis on semantics, which has a significant impact on viewers' sensitivity to video sequence quality.

2 Research Methodology

In this research, we simulated an environment of a real packet network and tested video sequences that would simulate the diverse content of teaching. We artificially degraded the quality of video sequences by packet loss and jitter. The objective of the test was to compare subjective methods with objective methods and evaluate the impact of the quality of the audio on the overall quality of the video sequence. We also wanted to show that semantics has a significant impact on viewers' sensitivity to the quality of the video sequence [6].

Subjective quality cannot be represented by an exact figure. Due to its inherent subjectivity, it can only be described statistically. Even in psychophysical threshold experiments, where the task of the observer is just to give a yes/no answer, there is a significant variation in contrast sensitivity functions and other

critical low-level visual parameters between 50 different video quality observers. When the artifacts become supra-threshold, the observers are bound to apply different weightings to each of them [18].

International recommendations for subjective methods of quality testing include specifications on how to implement different types of subjective tests. Some of these test methods are known as "double stimulus" methods where an observer evaluates quality or quality change between two (reference and test) video sequences. There are also "single stimulus" methods where the observer evaluates the quality of just one (test) video sequence [19, 20, 21]. The following subsections 2.1 to 2.3 describe three subjective methods: two “double stimulus” methods DSCQS and DSIS and one “single stimulus” ACR method. Subsections 2.4 and 2.5 introduce the metrics MSE and PSNR and SSIM index used in objective evaluation methods. Finally, subsection 2.6 presents the structural process model we created for evaluating video sequences.

2.1 DSCQS Method

The Double Stimulus Continuous Quality Scale (DSCQS) method is suitable for measuring the quality of the system that is related to the reference value as the observer is not familiar with the reference sequence order [19]. DSCQS is quite sensitive to small differences in quality and is thus the preferred method when the quality of the test sequence and reference sequence are similar [18].

2.2 DSIS Method

The Double Stimulus Impairment Scale (DSIS) method is suitable for assessing the extent of degradation of the test sequence as compared to the reference one, especially in case of visible/significant degradation. For example, it is used to evaluate the degradation of the sequence during transport. This method is faster than DSCQS since the sequences are displayed only once [19]. Subjects rate the amount of impairment in the test sequence on a discrete five-level scale ranging from “very annoying” to “imperceptible”. The DSIS method is well suited for evaluating clearly visible impairments such as artifacts caused by transmission errors [18].

2.3 ACR Method

The Absolute Category Rating (ACR) method is a single stimulus method; viewers only see the video under test, without the reference. They give one rating for its overall quality using a discrete five-level scale from “bad” to “excellent”. The fact that the reference is not shown with every test clip makes ACR a very efficient method compared to DSIS or DSCQS, which take almost 2 to 4 times longer, respectively [18, 20].

2.4 MSE and PSNR

The best-known methods for objective evaluation of signal quality include metrics based on pixel comparisons, such as MSE (Mean Squared Error) and PSNR (Peak Signal to Noise Ratio). An advantage of these methods is the speed and ease of calculation. A disadvantage is that they do not accurately capture the perception of quality and distortion by the human visual system [22].

The MSE is the mean of the squared differences between the gray-level values of pixels in two pictures or sequences *I* and *I'*:

$$\mathrm{MSE} = \frac{1}{\mathrm{TXY}} \sum_t \sum_x \sum_y \left[\mathrm{I}(t,x,y) - \tilde{\mathrm{I}}(t,x,y) \right]^2 \quad (1)$$

for pictures of size *X* x *Y* pixels and *T* frames in the sequence [22].

The PSNR in decibels is defined as:

$$\mathrm{PSNR} = 10 \log \frac{\mathrm{m}^2}{\mathrm{MSE}} \quad (2)$$

where *m* is the maximum value that a pixel can take [22].

2.5 SSIM Index

Newer methods for objective evaluation of the signal quality include the SSIM Index (Structural Similarity Index). The SSIM metric measures three components: the luminance similarity, the contrast similarity, and the structural similarity and combines them into one final value that determines the quality of the test sequence (Figure 1). This method differs from the above-described error-based methods described by using the structural distortion measurement instead of the error one [23]. It is due to the human visual system that is highly specialized in extracting structural information from the viewing field and it is not specialized in extracting the errors. Owing to this factor, the SSIM metric achieves a good correlation to subjective impression [24, 25]. The results are in the interval [0,1], where 0 and 1 denote the worst and the best quality, respectively.

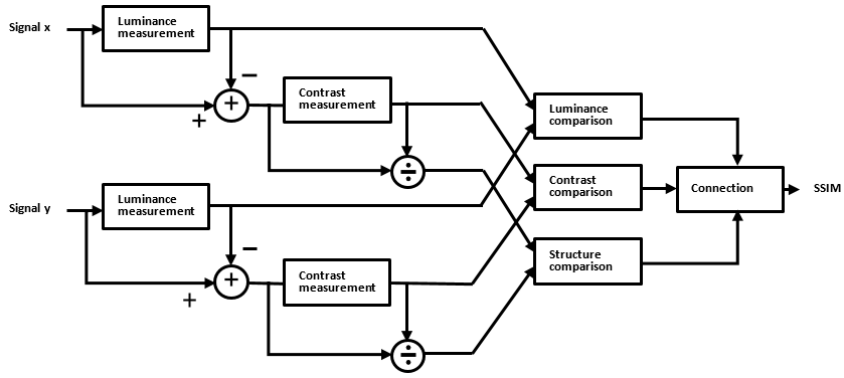

Figure 1
The block diagram of SSIM metric [26]

2.6 Process of Subjective Evaluation of Video Sequences

There is no single and ideal method to measure video quality. It is very important to choose the right method to meet our needs. Subjective methods provide more reliable results but objective methods are not influenced by the viewer's opinions or experiences [27]. Subjective video quality testing is difficult not only because of the time-consuming nature of testing itself but also due to the complexity of the steps that precede the actual testing. Figure 2 describes five steps of the process model we have designed for subjective evaluation of the quality of video sequences. It is based on the process model that we presented in the article [21].

Figure 2
The structural process model of video sequences quality evaluation [21]

2.6.1 Recording and Coding of Test Sequences

Reference video sequences were created based on real video calls. These sequences were recorded using the Logitech C270 web camera with HD resolution of 1270 x 720 pixels, utilizing the Logitech Webcam Software shipped with the web camera. Due to the purpose of the testing, it was important to create diverse demonstrations with a different emphasis on content, the importance of video or audio capture. Four types of reference video sequences are described below.

In the first test sequence (video sequence No. 1), the intention was to create a preview where the emphasis would be on the picture detail. The lecturer in this video preview informs students that if they have any questions, they can contact him at his e-mail address. The person in the preview does not pronounce this e-mail address but writes it on the board (Figure 3). So the only way this e-mail address information gets to the user of the videoconference is assuring that the image quality will be sufficient, to recognize it without difficulty.

Figure 3
Photo from the test video sequence No. 1

In the second and third test video sequences, the aim was to create a demonstration where an emphasis would be placed on the quality of the audio during static image transfer. In the second example (video sequence No. 2), a woman asks the recipient to contact someone by phone. She dictates her name and phone number. In the third test sequence (video sequence No. 3), the student asks a classmate to provide him with the lecture notes he missed. He uses several shortcuts, so passing the information takes a short time. Unlike in the first demonstration, in the video sequences two and three, the information is provided only in the form of sound. Therefore, to interpret it correctly, the audio must be captured completely and correctly.

In the fourth test sequence (video sequence No. 4), the teacher explains the formula for calculating electrical efficiency. The formula is written on the board, while the teacher simultaneously talks about individual variables in the formula. Since the information is provided through both image and sound at the same time, minor audio outages can be compensated for by the visual clarity of information or vice versa minor video outages can be compensated for by the audio clarity.

Each video sequence was encoded, because the video and audio formats used, as well as bit rates, do not match those used in videoconferencing. Recording and coding technical parameters of reference video sequences are described in Tab. 1.

Table 1
Recording and coding technical parameters

parameter	recording	coding
pixel	1270 x 720	1270 x 720
frame rate	15	30
sequence length [sec]	10	10
video format	WMV2	MPEG-4 AVC
audio format	WMA	AAC
bit rate of video [kbit/sec]	3535	1024
bit rate of audio [kbit/sec]	1411	128
audio sampling rate [kHz]	48	22.05

2.6.2 Degradation of Test Sequences

An important part of the research was the selection of appropriate subjective methods for evaluation of the quality of video sequences. As we wanted to use "double stimulus" methods in testing, we had to create degraded samples in addition to reference samples. To introduce degradations into the reference videoconferencing sequences, it was necessary to emulate the transfer environment through which the sequences were transmitted (Figure 4).

Network emulation is a process by which we can control and repeatedly simulate network performance. The changes in network parameters such as latency and

packet loss are provided by traffic shapers. They must be controlled according to predefined specifications to simulate the required features of the network.

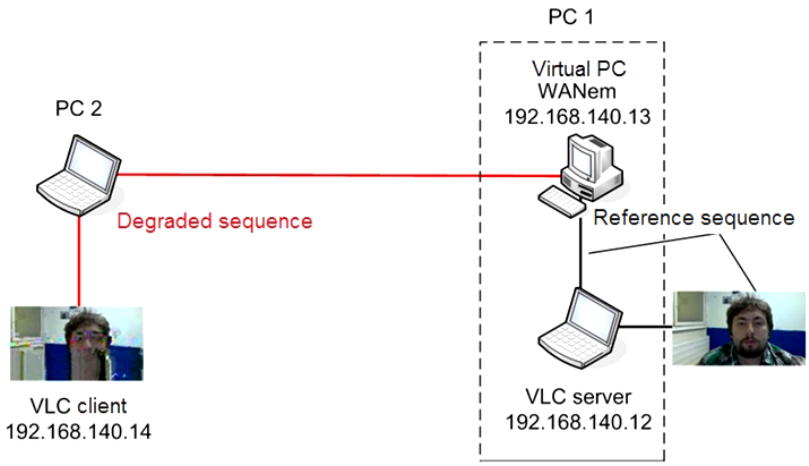

Figure 4
Network model for creation of degraded sequences [21]

PC 1 served as a video streaming server (Figure 4). We had to set the destination IP address, data transfer protocol (UDP), port, and modify the routing table to route all outgoing packets to the virtual PC. On the PC 2 side, VLC media player 0.8.6f was used as a client to receive the streamed video and also allowed to save it. Similarly, to the server, it was necessary to set the destination IP address (PC 2 IP address), data transfer protocol (UDP), port and address where the received video should be stored [21, 23].

Another program we used was WAN from TATA Consultancy Services (Figure 4). It is an open-source program used to emulate WAN networks (e.g. Internet) in a LAN environment. It allows setting many parameters such as bandwidth for transmission, latency, jitter and packet loss [28].

Each of the four reference samples was degraded by packet loss (0.5%, 1%, 3%, 5%, and 10%) and jitter (50 ms jitter at 100 ms latency).

2.6.3 Selection of Appropriate Methods

Absolute Category Rating (ACR) and Double Stimulus Impairment Scale (DSIS) methods were selected for the subjective evaluation of video samples. The ACR method has the advantage of being fast as the evaluator watches the sample only once and the length of the sample is relatively short (about 10 seconds). The DSIS method was also selected because of its time efficiency and the ability to capture more accurate differences between degraded samples, as we also have a reference sample for this method [20, 21]. The choice of suitable methods was also

influenced by the fact that both the ACR's and DSIS's outputs are MOS scores with values ranging from 1 to 5, so the results can easily be compared [14].

To objectively evaluate the quality of video sequences we used the MSU Video Quality Measurement Tool. From the portfolio of available methods, we chose the PSNR method, whose advantage is the speed of calculation [29]. The second objective method we used was the SSIM method that already includes models of the human visual system, and therefore, the results should better correspond to the outcomes of subjective evaluation [26]. Both methods required a comparison of the degraded video sequence with the reference sequence.

2.6.4 Preparation of Test Scenarios and Selection of Respondents

Since the testing was performed within the VLC multimedia player environment, it was necessary to create playlists in which the individual video sequences were arranged appropriately. To prepare the scenarios and the course of the subjective measurements, a program was created in the C# programming language. To play a video sequence the program uses an open-source DmediaPalyer that is a modification of the VLC player. The program consists of two parts: test manager part and tester part (Figure 5). The Test manager part is an interface used to create structure of the test. You can choose the type of subjective method, test sequence, reference sequence (if necessary) and enable or disable sound step-by-step. We presented this program in the article [27].

Figure 5
The block diagram of testing and test scenarios preparation program [27]

The ITU-T Recommendations specify that the number of respondents for subjective quality assessment must be greater than 4 and less than 40 [19, 20]. Based on this, we selected 20 respondents (10 women and 10 men), aged 20-51.

The fifth step of the subjective quality evaluation includes testing. The course of testing, evaluation, and comparison of the results are described in the following section.

3 Comparison of Video Sequences Quality Measurement Results

Our research aimed to compare subjective and objective methods of video sequence testing and to determine the degree of impact of audio quality on overall video quality with respect to the semantics.

Due to the time-consuming manual processing of results, two programs were created.

The first program was written in the C# programming language. There are two list data structures, one for each sequence. These data structures store individual objects whose variables have values read from individual result files. In the case of the DSIS method, the variables are the method name, respondent name, age, gender, reference and ranked sequence name, and the evaluation itself. In the case of the ACR method, the variables are the method name, respondent name, age, sex, names of the first and second sequence to be evaluated, and their evaluation itself. The program processes each file sequentially. After reading all the data, it checks whether the list contains an object with the same values of the variables. If there is no such object, the object with the loaded variables is saved. Otherwise, the object is deleted and a message about its deletion is written to the console. The algorithm then sequentially scans individual objects and writes them to the output file according to the given criteria. It also allows the results to be processed with respect to their statistical processing (performed by the second program described below). If the respondent was excluded from the DSIS method, they are also excluded from the ACR method.

The program has two outputs in the form of text files. In the first one, the results are processed according to the evaluation of the individual sequences. In the case of the DSIS method, the format is the reference sequence name and the test sequence name followed by five numbers. In the case of the ACR method, the format is the test sequence name and five numbers. The five numbers correspond to the evaluation scale of the given methods [20, 21]. If it has been chosen to take the statistical processing into account, the output is in the same file. In the second text file, the results are processed according to respondents who evaluated individual sequences. This output is needed for statistical processing of results for the DSIS method.

The second program is used for statistical processing of measured results. It was created in Matlab version R2008b. The algorithm for statistical processing of measured results was designed as follows:

The average score $\bar{u}_{jkr}$ is calculated for each test sequence

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \tag{3}$$

where u_{ijkr} is the respondent score i for test condition j, sequence k and number of repetitions r. N is the total number of respondents.

The standard deviation S_{jkr} and the peak coefficient β_{2jkr} are also calculated for each test sequence:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{4}$$

$$\beta_{2\,jkr} = \frac{m_4}{(m_2)^2} \tag{5}$$

where $$m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{6}$$

Then we find Q_i and P_i for each respondent i as follows:

If $2 \leq \beta_{2jkr} \leq 4$ then

$$\text{if } u_{ijkr} \geq \bar{u}_{jkr} + 2\,S_{jkr} \text{ then } P_i = P_i + 1 \tag{7}$$

$$\text{if } u_{ijkr} \leq \bar{u}_{jkr} - 2\,S_{jkr} \text{ then } Q_i = Q_i + 1 \tag{8}$$

If $\beta_{2jkr} < 2$ *or* $\beta_{2jkr} > 4$ then

$$\text{if } u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20}\,S_{jkr} \text{ then } P_i = P_i + 1 \tag{9}$$

$$\text{if } u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20}\,S_{jkr} \text{ then } Q_i = Q_i + 1 \tag{10}$$

The assessment of respondent i will not be taken into account if conditions (11) and (12) apply simultaneously:

$$\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05 \tag{11}$$

$$\left| \frac{P_i - Q_i}{P_i + Q_i} \right| < 0.3 \tag{12}$$

where J is the number of test conditions, K is the number of test sequences, and R is the number of repetitions.

The output of the program is a text file with the names of the individual respondents who were excluded based on the above algorithm.

The processing of data from objective evaluation methods consisted of a mathematical evaluation of each method for each test sequence. The MSU Video Quality Measurement Tool was used for this evaluation [29]. The program

supports a large number of video formats and objective methods and allows visualization of results or their subsequent saving in text form to a file.

Table 2 lists the summary of video sequences quality measurement results. In the case of subjective evaluation, the video sequences were rated by MOS scores that range from 1 to 5 [19, 20]. A video sequence rated by the score of 4 or higher is considered to be of high quality [9, 14].

The *ACR d.a.* and *DSIS d.a.* columns show the results for video sequences degraded by packet loss (0.5%, 1%, 3%, 5%, and 10%) and jitter (50 ms jitter in 100 ms latency). The *ACR r.a.* and *DSIS r.a.* columns show the results for video sequences degraded by packet loss and in which the degraded audio track was replaced by the audio track from the reference sequence. From the obtained results, it is clear that as the sequences deteriorate, the quality of the sequences decrease, both objectively and subjectively.

Comparing the evaluations for the ACR and DSIS methods, we found that video sequences were rated by a higher score when the DSIS method was used. This difference can be explained by the fact that in the case of the DSIS method the respondent was influenced by the reference sample.

Even at 0.5% and 1% packet loss degradation, some video sequences with the reference audio track received higher ratings than those with the original disturbed audio track. The results also imply that, in general, the degradation caused by jitter (50 ms jitter in 100 ms latency) does not affect the quality ratings as much as the degradation due to packet loss.

Table 2

Results of quality evaluation of test sequences (d.a. – degraded audio, r.a. – reference audio)

Video sequence	**Degradation**	**Subjective methods**				**Objective methods**	
		ACR d.a.	**ACR r.a.**	**DSIS d.a.**	**DSIS r.a.**	**PSNR**	**SSIM**
No. 1	**Packet loss 0.5%**	4.70	4.65	4.80	4.75	42.75	0.98
	Packet loss 1%	3.55	3.90	4.15	3.90	35.34	0.97
	Packet loss 3%	2.25	3.15	3.15	3.30	30.52	0.92
	Packet loss 5%	2.15	2.65	1.95	2.85	28.03	0.86
	Packet loss 10%	1.00	1.95	1.05	2.35	25.65	0.84
	Latency 100 ms, Jitter 50 ms	3.65	x	3.75	x	44.14	0.98
No. 2	**Packet loss 0.5%**	2.45	3.25	3.25	3.70	33.41	0.95
	Packet loss 1%	2.10	3.40	2.30	3.80	33.31	0.95
	Packet loss 3%	1.55	2.65	2.05	3.20	31.15	0.91
	Packet loss 5%	1.70	2.20	1.35	2.25	24.00	0.86
	Packet loss 10%	1.10	1.70	1.00	1.85	18.99	0.78

	Latency 100 ms, Jitter 50 ms	4.15	x	3.85	x	43.14	0.98
No. 3	**Packet loss 0.5%**	3.50	3.05	4.45	4.30	34.57	0.96
	Packet loss 1%	2.85	3.20	3.65	3.60	26.89	0.87
	Packet loss 3%	1.25	2.80	2.30	3.45	25.52	0.84
	Packet loss 5%	1.20	2.75	1.35	2.65	22.97	0.77
	Packet loss 10%	1.00	2.05	1.00	1.95	18.38	0.66
	Latency 100 ms, Jitter 50 ms	3.50	x	3.85	x	35.56	0.96
No. 4	**Packet loss 0.5%**	4.20	4.25	4.15	4.05	31.03	0.95
	Packet loss 1%	3.65	3.75	3.95	3.85	30.78	0.94
	Packet loss 3%	2.25	2.50	2.10	2.65	23.06	0.84
	Packet loss 5%	1.45	2.10	1.45	2.35	20.32	0.79
	Packet loss 10%	1.00	2.00	1.00	1.70	17.40	0.74
	Latency 100 ms, Jitter 50 ms	4.35	x	3.95	x	43.99	0.98

As we have assumed, the subjective evaluation was also influenced by pictorial information. From Figure 6, showing the comparison of the evaluation of the video sequences by the ACR method, we can clearly see that the video sequences No. 2 and No. 3 were evaluated by the lowest marks. In these video sequences, the image being transmitted was static and an emphasis was placed on the content of the audio. In the case of the objective assessment (Table 2), this difference has not been proved to such an extent.

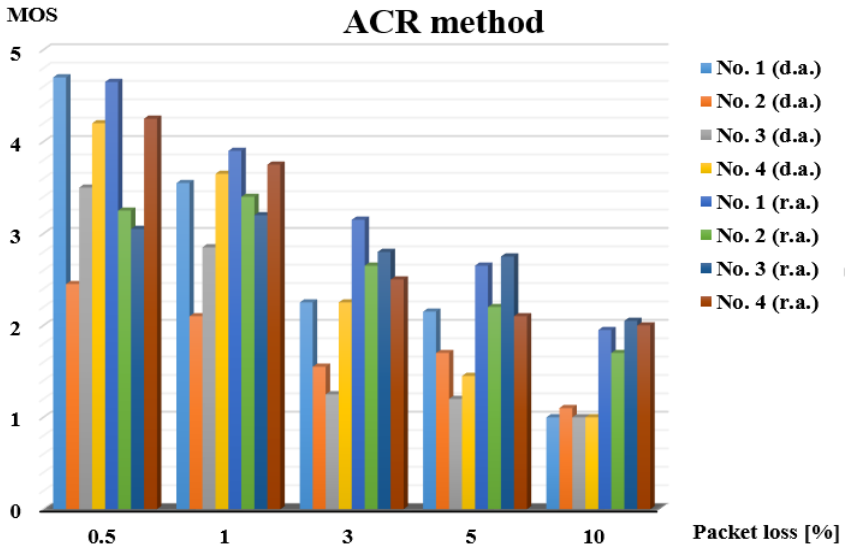

Figure 6
Comparison of the evaluation of the video sequences by the ACR method

The results of the subjective quality evaluation have shown that under the ideal conditions in the transmission network (without packet loss and latency) the quality of videoconferencing has been rated as "good" (MOS > 4). Therefore, from the perspective of the user, the video frame resolution, audio and video bitrate, and the used codecs provide the user with sufficient quality.

However, each internet protocol (IP) based transmission network will cause packet loss and latency. Their source is the non-link structure of the network. Quality codecs can at least partially compensate for the loss of information transmitted [23]. The results of the subjective quality assessment of various distorted video sequences have confirmed that packet loss of less than 1% must be achieved to obtain a very good quality videoconference.

In subjective methods (ACR, DSIS), the lowest score was evaluated for sound-related sequences (No. 2, No. 3). This confirmed that both the content of the information transmitted and the clarity of information for the evaluator play an important role in subjective quality assessment. Of course, in the videoconference that supports the learning process, the other receiving party must at least partially understand the lecture or lesson issues.

Figure 7 shows a comparison of the evaluation of the video sequences by the SSIM method. The resultant SSIM index is a decimal value between -1 and 1. The value of 1 is only reachable in the case of two identical sets of data and therefore indicates perfect structural similarity. A value of 0 indicates no structural similarity [22].

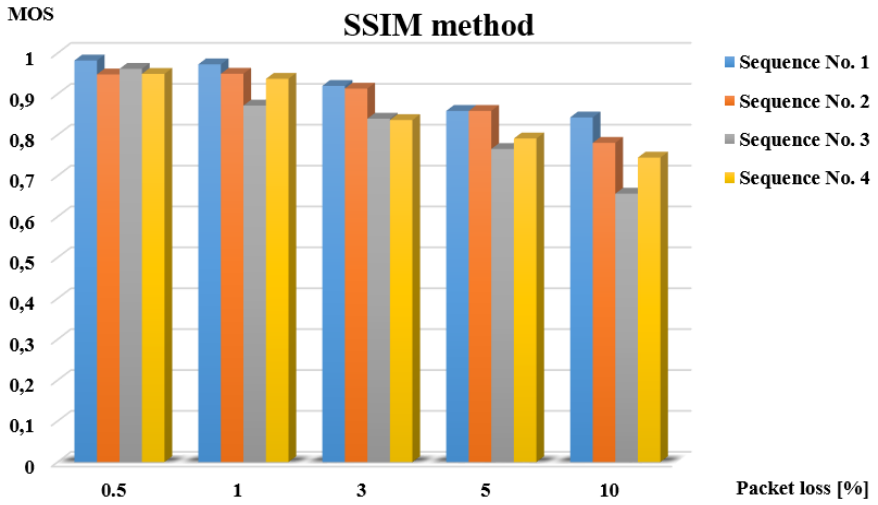

Figure 7
Comparison of the evaluation of the video sequences by the SSIM method

When evaluating video sequences using objective methods (PSNR, SSIM), video sequences No. 3 and No. 4 were scored by the lowest marks (Figure 7, Table 2).

So, we can conclude that the results of subjective and objective methods are different for our research samples. This implies that we still do not have objective methods available to replace demanding and lengthy subjective evaluation.

Based on the results of the subjective evaluation of the sequences with the original audio and the sequences in which the degraded audio was replaced by the reference, we see that for the packet loss of 3% and 5% the sequences with the reference audio are rated much higher (often by more than 1 point on the MOS scale). The difference between individual sequence evaluations is much smaller in samples with the reference audio compared to sequences with the original audio track. In our research, we have confirmed that the quality of audio has a great impact on the overall quality of videoconferencing. In future work, we can investigate whether a similar trend is observed when changing the tasks, that is, if we gradually insert different deteriorated audio tracks into the reference video sequence.

From the measured values, it also follows that in the case of 10% packet loss the respondents rated with the worst possible marks ("bad" or "poor"). In future research, the degradation with packet loss of over 10% would not make sense to test. However, it would be interesting to extend the tests with a greater number of sequences or more types of deterioration. A significant disadvantage of subjective tests is that they are time-consuming, which to a large extent limits their use. With a higher number of test sequences or a higher number of evaluators, we no longer recommend using a questionnaire for writing but a suitable software tool that would also facilitate the evaluation process.

3.1 Correlation between Objective and Subjective Methods

The correlation coefficient describes the direction and the magnitude of the relationship between two variables. It is calculated as follows:

$$r_{yx} = \frac{k_{xy}}{\sigma_x \sigma_y} \tag{13}$$

where σ_x a σ_y are standard deviations of variables x and y, respectively, and k_{xy} is their covariance calculated as:

$$k_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

The value of a correlation coefficient ranges between −1 and 1. The greater the absolute value of a correlation coefficient, the stronger the linear relationship. The strongest linear relationship is indicated by a correlation coefficient of −1 or 1. The weakest linear relationship is indicated by a correlation coefficient equal to 0. A positive correlation means that if one variable gets bigger, the other variable

tends to get bigger. A negative correlation means that if one variable gets bigger, the other variable tends to get smaller [26].

For each test sequence, the correlation coefficients between particular objective (SSIM, PSNR) and subjective (ACR, DSIS) methods were calculated (Table 3).

Table 3
Correlation between objective and subjective methods

	ACR d.a.	**ACR r.a.**	**DSIS d.a.**	**DSIS r.a.**
SSIM	0.834	0.798	0.858	0.881
PSNR	0.853	0.850	0.827	0.927

The results show that the highest correlation is between the objective metric PSNR and the subjective method DSIS with reference audio (Table 3). However, correlation results cannot be generalized based on our measurements. In general, there is no objective method by which we can completely replace the subjective perception of a person.

Conclusions

Videoconferencing technology brings vast new possibilities into the process of modern education and overcomes distance barriers. Combined with interactive computing technology, it represents the technology of the future, in the learning process.

Increasing transmission speeds in today's modern networks enable us to provide new e-learning support services such as videoconferencing, on-demand streaming, or online streaming. Both voice services (VoIP) and moving image transfer services need to be monitored to see if the service is of adequate quality to the customer. This quality monitoring must necessarily be automated because it would be impractical, financially demanding and vulnerable to errors, to employ people for these activities.

This experiment compared the subjective methods of evaluating videoconferencing quality with known objective methods and thereby contribute to the development of new objective metrics. Time-consuming subjective measurements were supported by models and programs that simplified scenario preparation, testing and results processing. These will be used in further research dealing with the measurement of video sequences quality.

The results of our comparison have confirmed that we still do not have an objective method that can fully substitute the time-consuming subjective testing. Based on the results of the subjective evaluation of sequences, with the original audio track and the sequences in which the degraded audio track was replaced by the audio track from the reference sequence, we have confirmed that the quality of the audio has a significant impact on the overall quality of videoconferencing and the ultimate understanding of its content. As a result, if any video information is

supported by relevant audio information, we can compensate for the loss of video information by improving the audio quality. We can also influence the quality of videoconferencing by ensuring correct pronunciation, intelligibility and articulation.

References

[1] Azimi-Sadjadi, B. et al.: Robust Key Generation from Signal Envelopes in Wireless Networks. In CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security, pp. 401-410, New York, NY, USA, 2007

[2] Bisták P., et al.: Utilisation of Videoconferencing for Education. 1st ed., Elfa Kosice, 2005

[3] Haffner, et al.: The multimedia as a form of modern education. In QUAERE 2018, Hradec Králové: Magnanimitas, pp. 898-907, 2018

[4] Haffner, et al.: Multimedia support for education of mechatronics. In 2018 Cybernetics & Informatics (K&I): 29th International Conference. Lazy pod Makytou, Slovakia, 2018

[5] Whittaker, S.: Video as a technology for interpersonal communications: a new perspective. Proc. SPIE 2417, Multimedia Computing and Networking, 1995, https://doi.org/10.1117/12.206055

[6] Frater, M. R., Arnold, J. F., & Vahedian, A.: Impact of audio on subjective assessment of video quality in videoconferencing applications, In *IEEE* Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 9, pp. 1059-1062, Sept. 2001

[7] Heribanová, P., Polec, J., Poctavek, J. & Mordelová, A.: Intelligibility Threshold for Cued Speech in H.264 Videoconference. International Journal of Electronics and Telecommunications, 57, pp. 383-387, 2011

[8] Coudoux, F.-X., Gazalet, M. G., Derviaux, C. & Corlay, P.: Picture quality measurement based on block visibility in discrete cosine transform coded video sequences. Journal of Electronic Imaging. 2001, https://doi.org/10.1117/1.1344184

[9] Wang, Z. & Bovik, A.: Handbook of video and image processing. Academic Press. 2000

[10] Káňa, L.: Elektroakustika, ČVUT Brno, Česká republika, 2013

[11] Saadatzi, M., Saadatzi, M. N., Tavaf, V. & Banerjee, S.: Development of a PVDF based artificial basilar membrane. Proc. SPIE 10593, Bioinspiration, Biomimetics, and Bioreplication VIII, 2018

[12] Andriichenko, O. O., Denysenko, O. I.: Subjective evaluation of the clarity of the noisy language in the lecture room, Electronic and Acoustic Engineering, 2, 3, 55-60, 2019

[13] Takács, A. et al: Models for Force Control in Telesurgical Robot Systems. Acta Polytechnica Hungarica, 12, pp. 95-114, 2015

[14] Rizek, H., Brunnström, K., Wang, K., Andrén, B. & Johanson, M.: Subjective evaluation of a 3D videoconferencing system. Proc. SPIE 9011, Stereoscopic Displays and Applications XXV, 2014

[15] Mardiak, M. & Polec, J.: Novel method for objectively measuring video quality. Proceedings ELMAR-2010, Zadar, pp. 109-112, 2010

[16] Zaric, A. et al.: Image quality assessment - comparison of objective measures with results of subjective test. Proceedings ELMAR-2010, Zadar, pp. 113-118, 2010

[17] Wang, Z. & Bovic, A. C.: A universal image quality index. IEEE Signal Processing Letters, Vol. 9, pp. 81-84, March 2002

[18] Winkler, S.: Digital video quality vision model and metrics. Chichester: John Wiley & Sons Ltd, 2005

[19] ITU-R Recommendation ITU-R BT.500-11. Methodology for the subjective assessment of the quality of television pictures, 2002

[20] ITU-T Recommendation ITU-T P.910. Subjective video quality assessment methods for multimedia applications, 2008

[21] Filanová, J. & Mardiak, M.: Meranie kvality video signálu. Elektrorevue, 15, pp. 32.1-6, 2010

[22] Wang, Z., Lu, L. & Bovic, A. C.: Video quality assessment using structural distortion measurement. Signal Processing: Image Communication, special issue on "Objective video quality metrics", Vol. 19, No. 2, pp. 121-132, 2004

[23] Votruba, A. & Medvecký, M.: Evaluation of the Effectiveness of QoS Provisioning in Ethernet Networks. EE časopis pre elektrotechniku a energetiku, 17, pp. 92-96, 2011

[24] Wu, H. R. & Rao, K. R.: Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications) Boca Raton: CRC Press, 2006

[25] Wu, H. R. & Rao, K. R. & Kassim, A.: Digital Video Image Quality and Perceptual Coding. Journal of Electronic Imaging 16(3), 2007

[26] Uhrina, M. & Hlubik, J. & Vaculík, M.: Correlation between Objective and Subjective Methods Used for Video Quality Evaluation. Advances in Electrical and Electronic Engineering. 11, 2012

[27] Mardiak, M. & Filanová, J.: Quality of a Video Signal. In: New Information and Multimedia Technologies. NIMT - 2008 : Brno, Czech Republic, 18.-19.9.2008. - Brno : Brno University of Technology, 2008

[28] Nambiar, M. et al.: WANem - Open Source software, Performance Engineering Research Centre, TATA Consultancy Services, Mumbai India, 2008

[29] Vatolin, D. et al.: MSU Quality Measurement Tool: Metrics information. Available: http://compression.ru/video/quality_measure/info_en.html